Πανεπιστήμιο Αιγαίου

Ανάλυση Κατηγορικών Δεδομένων

Ενότητα 10: Γενικευμένα γραμμικά μοντέλα

Στέλιος Ζήμερας

Τμήμα Μαθηματικών

Εισαγωγική Κατεύθυνση: Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών

Σάμος, Ιούνιος 2015

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Μοντέλα με μορφή δυαδική επιτυχία ή αποτυχία

$$P(Y_i = 0) = 1 - \pi_i \; ; \quad P(Y_i = 1) = \pi_i$$

Η συνάρτηση πιθανότητας της τυχαία μεταβλητή Y που ορίσαμε παραπάνω είναι : $f(y;\pi) = P(Y = y) = \pi^{y}(1-\pi)^{1-y}$.

Το π είναι η παράμετρος της συνάρτησης.

Θεωρούμε n τυχαίες δίτιμες μεταβλητές, Y_1 , Y_2 , … , Y_n οι οποίες είναι ανεξάρτητες με $P(Y_i = 1) = \pi_i$

Η **από κοινού πιθανότητα** είναι :

$$f(\pi; y) = \prod_{j=1}^{n} \pi_j^{y_j} (1-\pi_j)^{1-y_j} = \exp\left[\sum_{j=1}^{n} y_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j) \right]$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Όπου $\pi = [\pi_1, ..., \pi_n]^T$ και $y = [y_1 ... y_n]^T$, και ανήκει στην εκθετική οικογένεια κατανομών.

- **Ορισμός** : Λέμε ότι η κατανομή μιας τ.μ. Υ ανήκει στην Εκθετική Οικογένεια κατανομών όταν μπορεί να γραφτεί στη μορφή

$$f(y;\theta) = \exp[\sum b_i(\theta)\mathrm{T}_i(y) + c(\theta) + h(y)]$$

όπου h c T_i b_i θεωρούνται γνωστές συναρτήσεις.

απλούστερη μορφή:

$$f(y;\theta) = \exp[b(\theta)a(y) + c(\theta) + h(y)] \ .$$

$$f(y;\theta) = \exp[b(\theta)\, y + c(\theta) + h(y)]$$ **κανονική μορφή**

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

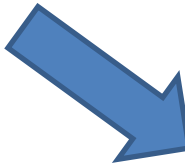
Στην περίπτωση που τα π_j είναι όλα ίσα μεταξύ τους μπορούμε να ορίσουμε τη συνάρτηση $Y = \sum_{j=1}^{n} Y_j$ στην οποία το Y παριστάνει τον αριθμό των «επιτυχιών» σε n ανεξάρτητες προσπάθειες και λέμε ότι το Y ακολουθεί τη διωνυμική κατανομή.

Εάν η τυχαία μεταβλητή Y έχει τη *διωνυμική κατανομή* τότε συμβολίζουμε με :

$Y \sim b(n,\pi)$ και έχει συνάρτηση πιθανότητας :

$$P(Y=y) = \binom{n}{y} \pi^{y} (1-\pi)^{n-y} \quad \text{όπου} \quad y = 0, 1, \ldots, n$$

$$Y_i \sim b(n_i, \pi_i)$$



$$l(\pi_1,\ldots,\pi_N; y_1,\ldots,y_N) = \sum_{i=1}^{N} \left\{ y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right\}$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Εστω N ανεξάρτητες μεταβλητές Y_i για τις οποίες ισχύει : $Y_i \sim b(n_i, \pi_i)$.

Με $n_1, n_2, \ldots, n_N$ συμβολίζουμε τον αριθμό δοκιμών που υπάρχουν σε κάθε ομάδα που περιγράφεται από μία μεταβλητή Y_i.

- Αναλογία των επιτυχιών: $P_i = y_i / n_i$ $\Rightarrow$ $0 \le \frac{y_i}{n_i} \le 1$.

- Μοντελοποίηση των πιθανοτήτων ως:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$E(Y_i) = \mu_i$$

$$g(\pi_i) = \eta_i = x_i^T \beta$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Στην περίπτωση αυτή για να συσχετίσουμε την πιθανότητα π_i με τη γραμμική έκφραση

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$$

πρέπει να χρησιμοποιήσουμε ένα γραμμικό μετασχηματισμό g(π) ο οποίος απεικονίζει το διάστημα (0,1) σε όλη την ευθεία (-∞, ∞). Υπάρχει μια μεγάλη ποικιλία από τέτοιες συναρτήσεις σύνδεσης.

Παράδειγμα

Ας θεωρήσουμε μία γλώσσα, η οποία προέρχεται από μία άλλη γλώσσα. Ένα απλό μοντέλο για την αλλαγή στο λεξιλόγιο είναι, αν οι γλώσσες, μεταβάλλονται σε χρόνο t, τότε η πιθανότητα να εμφανίζουν συγγενικές λέξεις, για μία συγκεκριμένη έννοια, είναι $e^{-\theta t}$, όπου θ είναι μία παράμετρος, κατά προσέγγιση η ίδια για πολλές κοινές έννοιες. Σε μία μελέτη N διαφορετικών εννοιών, υποθέτουμε ότι ένας γλωσσολόγος κρίνει αν οι λέξεις των δύο γλωσσών έχουν συγγένεια ή όχι για κάποια έννοια [2].

$$Y_i = \begin{cases} 1 & \text{αν οι δύο γλώσσες εμφανίζουν συγγενικές λέξεις για μία έννοια i} \\ 0 & \text{αν οι λέξεις δεν σχετίζονται} \end{cases} .$$

$$P(Y_i = 1) = e^{-\theta t}, \theta \geq 0 \qquad P(Y_i = 0) = 1 - e^{-\theta t}, \theta \geq 0$$

$$\theta = [0, \infty]$$

Παράδειγμα

Αυτή είναι μία περίπτωση διωνυμικής κατανομής (n, π) με $n = 1$ και $E(Y_i) = \pi = e^{-\theta t}$. Στην περίπτωση αυτή ως συνάρτηση σύνδεσης g χρησιμοποιήσαμε τη λογαριθμική:

$$g(\pi) = log(\pi) = -\theta t \ ,$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

- Τρεις όμως είναι αυτές που χρησιμοποιούνται στην πράξη.

1. Η λογιστική συνάρτηση $g_1(\pi) = \log \frac{\pi}{1-\pi}$

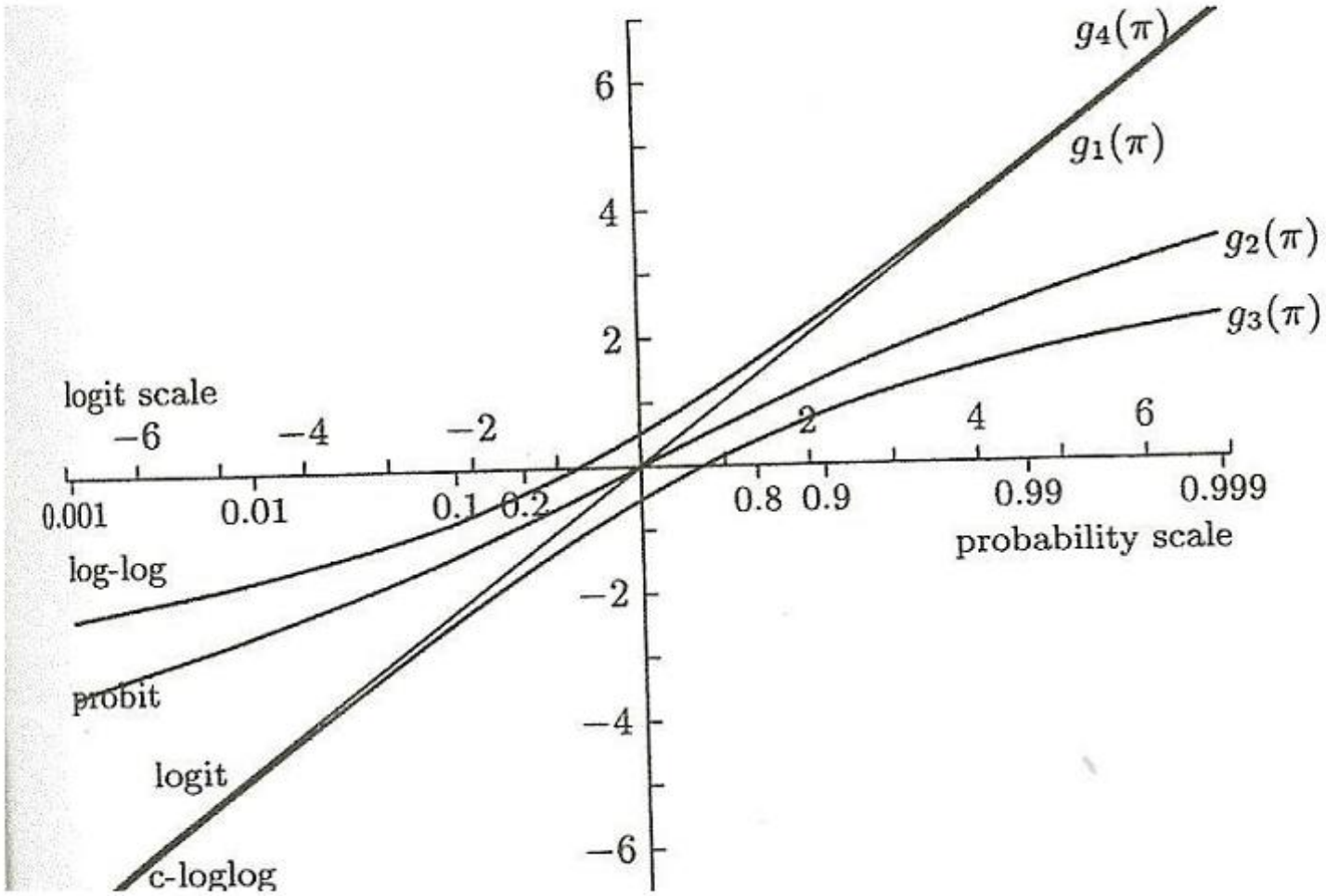
2. Η probit ή αντίστροφη κανονική συνάρτηση

$$g_2(\pi) = \Phi^{-1}(\pi),$$

3. η συμπληρωματική log-log συνάρτηση

$$g_3(\pi) = \log\{-\log(1-\pi)\}.$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ



ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Η logit και η Probit σχετίζονται σχεδόν γραμμικά για τιμές του π στο διάστημα $0,1 \leq \pi \leq 0,9$. Για τον λόγο αυτό είναι δύσκολη η διάκριση μεταξύ των δυο αυτών συναρτήσεων όταν πρόκειται για ζητήματα καλής προσαρμογής.

Για μικρές τιμές του π, η συμπληρωματική log-log συνάρτηση είναι κοντά στην λογιστική συνάρτηση

Όταν π τείνει στο 1 τότε η συμπληρωματική log-log συνάρτηση τείνει στο άπειρο πολύ πιο αργά σε σύγκριση με τις άλλες τρεις συναρτήσεις

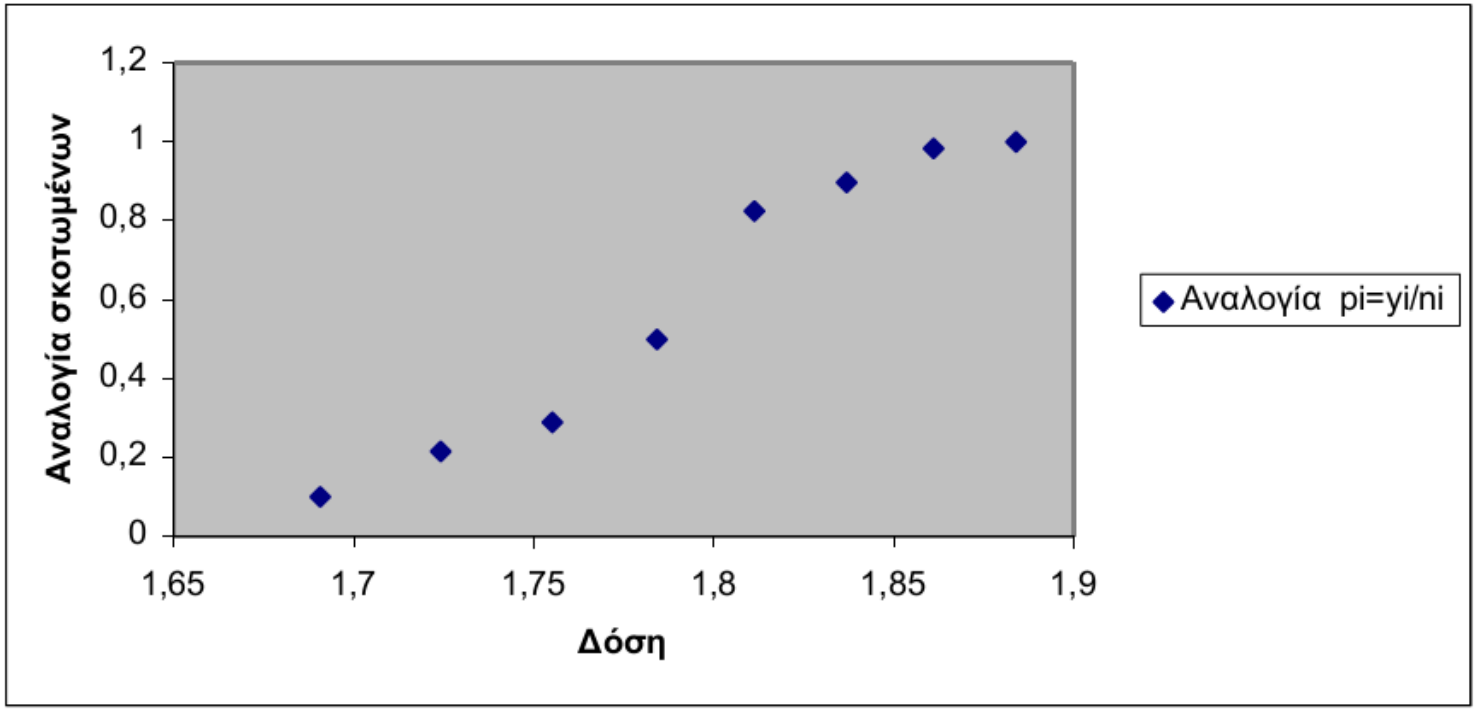
Παρομοίως η πιο αργή συνάρτηση στην περιοχή του0 είναι ηlog-log.

ΠΑΡΑΔΕΙΓΜΑ

Ο παρακάτω πίνακας δείχνει τον αριθμό εντόμων που πέθαναν μετά από έκθεση πέντε ωρών σε αεριώδη ανθρακικό ντιζουλφίδιο σε ποικίλες συγκεντρώσεις

Δόση x_i **($\log_{10}CS_2 mgl^{-1}$)**	**Αριθμός των** Εντόμων n_i	**Αριθμός που** Σκοτώθηκαν y_i	**Αναλογία** $p_i=y_i/n_i$
1, 6907	59	06	0, 102
1, 7242	60	13	0, 217
1, 7552	62	18	0, 290
1, 7842	56	28	0, 500
1, 8113	63	52	0, 825
1, 8369	59	53	0, 898
1, 8610	62	61	0, 984
1, 8839	60	60	1, 000

ΠΑΡΑΔΕΙΓΜΑ



Αναλογίες πιθανότητας

$$P_i = y_i / n_i$$

ΠΑΡΑΔΕΙΓΜΑ

ΠΡΟΣΑΡΜΟΓΗ ΛΟΓΙΣΤΙΚΟΥ ΜΟΝΤΕΛΟΥ

Περιγραφή της πιθανότητας επιτυχίας με $g(\pi) = \beta_1 + \beta_2 x$.

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

Οπότε:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i \quad \Rightarrow \quad \log(1-\pi_i) = -\log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right]$$

λογαριθμική συνάρτηση πιθανοφάνειας

$$l(\pi_1,\ldots,\pi_N; y_1,\ldots,y_N) = \sum_{i=1}^{N}\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i}\right\}$$

$$l = \sum_{i=1}^{N}\left[y_i(\beta_1 + \beta_2 x_i) - n_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\binom{n_i}{y_i}\right]$$

ΠΑΡΑΔΕΙΓΜΑ

Τα scores ως προς τα β_1 και β_2 θα είναι:

$$U_1 = \frac{\partial l}{\partial \beta_1} = \sum \left\{ y_i - n_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\} = \sum (y_i - n_i \pi_i)$$

$$U_2 = \frac{\partial l}{\partial \beta_2} = \sum \left\{ y_i x_i - n_i x_i \left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right] \right\} = \sum x_i (y_i - n_i \pi_i)$$

Πίνακας πληροφορίας

$$J = \begin{pmatrix} \sum n_i \pi_i (1 - \pi_i) & \sum n_i x_i \pi_i (1 - \pi_i) \\ \sum n_i x_i \pi_i (1 - \pi_i) & \sum n_i x_i^2 \pi_i (1 - \pi_i) \end{pmatrix}$$

ΠΑΡΑΔΕΙΓΜΑ

- **Scores**

$$\mathbf{u}(\boldsymbol{\theta}) = \begin{pmatrix} u_1(\boldsymbol{\theta}) \\ \vdots \\ u_p(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \dfrac{\partial l(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix}$$

- **Πίνακας πληροφορίας**

$$I(\boldsymbol{\theta}) = E_\theta\left[-H(\boldsymbol{\theta})\right] = E_\theta\left[-\frac{\partial U_j(\boldsymbol{\theta})}{\partial \theta_i}\right] = E_\theta\left[\frac{\partial^2 \log f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]$$

ΠΑΡΑΔΕΙΓΜΑ

Η στατιστική συνάρτηση πηλίκου λογαριθμικής πιθανοφάνειας είναι :

$$D = \sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n - y_i}{n - \hat{y}_i}\right) \right]$$

Εκτιμώμενες τιμές

Πίνακας 24: Προσαρμόζοντας το λογιστικό μοντέλο στα δεδομένα θανάτου σκαθαριών

	Αρχική Εκτίμηση	Πρώτη Προσέγγιση	Δεύτερη Προσέγγιση	Τέταρτη Προσέγγιση	Δέκατη Προσέγγιση
b_1	0	-37, 849	-53, 851	-60, 700	-60, 717
b_2	0	21, 334	30, 382	34, 261	34, 270

$D = 11, 23$

Εκτιμήσεις β_1 και β_1

Επειδή το πάνω 5% σημείο της x_6^2 κατανομής είναι 12, 59 συμπεραίνουμε ότι το μοντέλο δεν προσαρμόζεται ιδιαίτερα καλά στα δεδομένα.

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Για το γενικό λογιστικό μοντέλο, οι εκτιμήσεις μέγιστης πιθανοφάνει ας των παραμέτρων β και συνεπώς των πιθανοτήτων $\pi_i = g^{-1}\left(x_i^T \beta\right)$ προκύπτουν μεγιστοποιώντας τη συνάρτηση λογαριθμικής πιθανοφάνειας

$$l(\pi; y) = \sum_{i=1}^{N} \left[y_i \log \pi_i + (n_i - y_i) \log (1 - \pi_i) + \log \binom{n_i}{y_i} \right] .$$

Για να μετρήσουμε την «καλή προσαρμογή» του μοντέλου, χρησιμοποιούμε τη στατιστική συνάρτηση απόκλισης :

$$D = 2\left[l\left(\hat{\pi}_{\max}; y\right) - l\left(\hat{\pi}; y\right) \right]$$

όπου $\hat{\pi}_{\max}$ είναι το διάνυσμα των εκτιμήσεων μέγιστης πιθανοφάνειας που αντιστοιχεί στο πλήρες μοντέλο και $\hat{\pi}$ είναι το διάνυσμα των εκτιμήσεων μέγιστης πιθανοφάνειας του μοντέλου που μας ενδιαφέρει.

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Για το πλήρες μοντέλο παίρνουμε τα π_i ως τις παραμέτρους που θέλουμε να εκτιμήσουμε

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{n_i - y_i}{1-\pi_i} \quad \Rightarrow \quad y_i / n_i$$

$$l\left(\hat{\pi}_{\max}; y\right) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i}\right) + \left(n_i - y_i\right)\log\left(1 - \frac{y_i}{n_i}\right) + \log\binom{n_i}{y_i}\right]$$

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + \left(n_i - y_i\right)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Οπότε το D θα έχει τη μορφή : $D = 2\sum o \log \frac{o}{e}$

Όπου το ***o*** παριστάνει τις παρατηρηθείσες συχνότητες y_i και $(n_i - y_i)$ και το ***e*** παριστάνει τις αντίστοιχες εκτιμηθείσες αναμενόμενες συχνότητες ή προσαρμοσμένες τιμές $n_i\hat{\pi}_i$ και $\left(n_i - n_i\hat{\pi}_i\right)$.

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Για να ελέγξουμε την καλή προσαρμογή του μοντέλου, χρησιμοποιούμε την στατιστική συνάρτηση deviance:

$$D = 2[l(\hat{\pi}_{max}; y) - l(\hat{\pi}; y)].$$

Αν οι μεταβλητές απόκρισης $Y_1, ...Y_N$ είναι ανεξάρτητες και ακολουθούν τη διωνυμική κατανομή,τότε, η λογαριθμική συνάρτηση πιθανοφάνειας, είναι:

$$l(\mathbf{b}, \mathbf{y}) = \sum_{i=1}^{T} \{y_i log\pi_i - y_i log(1-\pi_i) + n_i log(1-\pi_i) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}\}.$$

$$\hat{\pi}_i = y_i/n_i$$

$$l(\mathbf{b}_{max}; \mathbf{y}) = \sum [y_i log \begin{pmatrix} y_i \\ n_i \end{pmatrix}] - y_i log(\tfrac{n_i - y_i}{n_i}) + n_i log(\tfrac{n_i - y_i}{n_i}) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}.$$

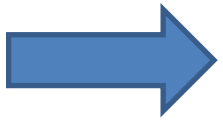
$$l(\mathbf{b}; \mathbf{y}) = \sum [y_i log \begin{pmatrix} \hat{y}_i \\ n_i \end{pmatrix}] - y_i log(\tfrac{n_i - \hat{y}_i}{n_i}) + n_i log(\tfrac{n_i - \hat{y}_i}{n_i}) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix},$$

$$\hat{y}_i = n_i \hat{\pi}_i$$

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

$$D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})]$$

$$= 2\sum_{i=1}^{N}\{y_i log\left(\begin{array}{c} y_i \\ \hat{y}_i \end{array}\right) + (n_i - y_i) log(\tfrac{n_i - y_i}{n_i - \hat{y}_i})\},$$

$$D = 2\sum olog\tfrac{o}{e},$$  $$D \sim \chi^2_{N-p},$$